INTRODUCTION TO DATA SCIENCE

CAPSTONE PROJECT

PREDICTING STOCK PRICES WITH GAUSSIAN PROCESS

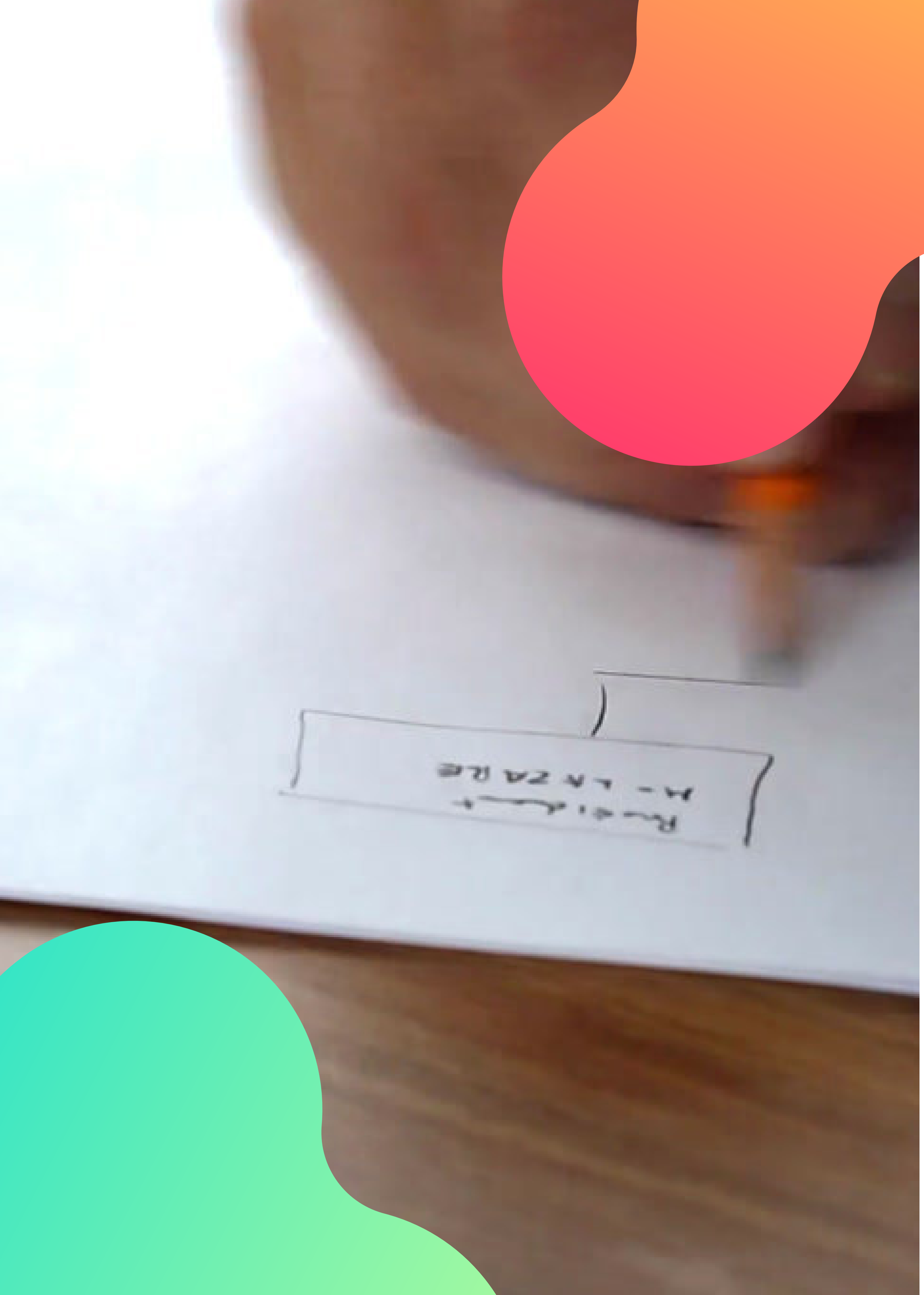
Nguyễn Viết Mạnh Khoa - 20184278
Phí Hoàng Long - 20184288
Lê Minh Hiếu - 20184257

OUTLINE

1. Introduction
2. Data crawling
3. Data visualization
4. Algorithm
5. Evaluation

Work Contribution

- Le Minh Hieu: crawling data and preprocessing data
- Phi Hoang Long: visualize data and give insight observations
- Nguyen Viet Manh Khoa: implementing Gaussian Process method to predict stock price and other methods for comparison purpose.



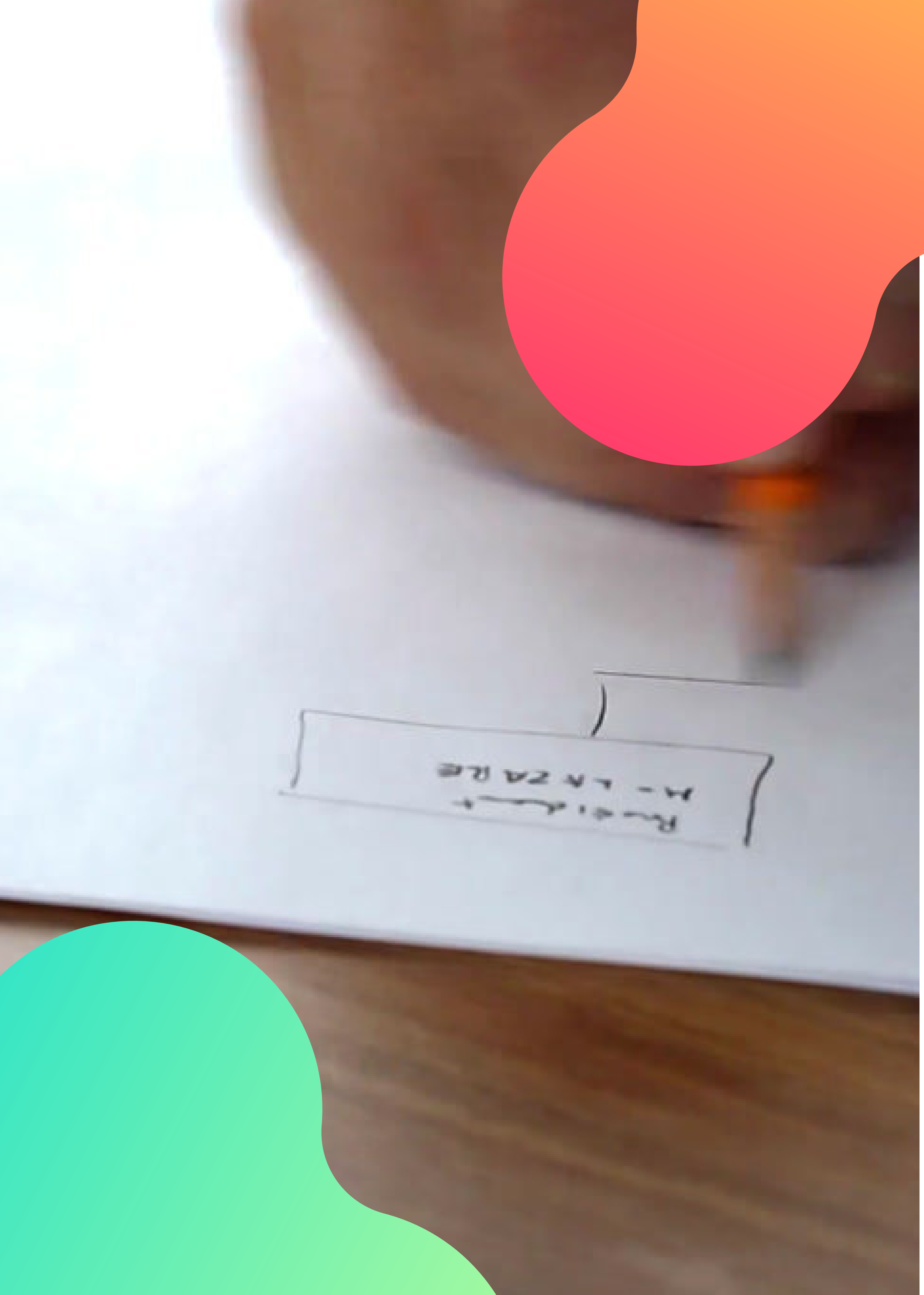
INTRODUCTION

Introduction - Stock price prediction

- A classic and important problem.
- Gain insight about market behavior over time
- Spotting trends hard to predict by human means
- Machine learning is an efficient method

Introduction - Learning problem

- Task: predict and gain future stock price information
- Experience: past stock price trend pattern
- Performance: accuracy with ground truth



DATA CRAWLING

Yahoo Finance

1. NASDAQ Composite (^IXIC)

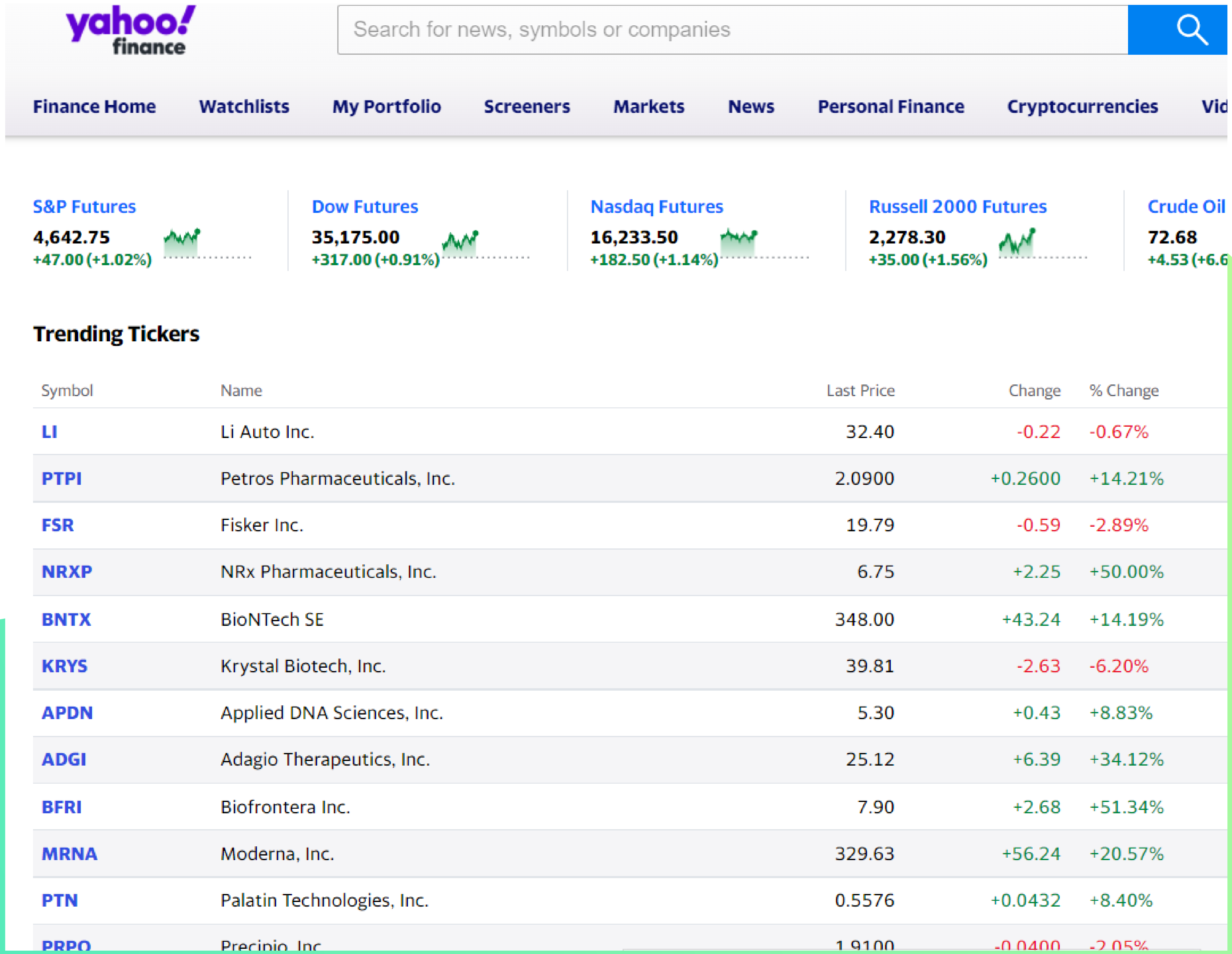
Includes almost all stocks on NASD

2. Palatin Technologies (PTN)

A biopharmaceutical company

3. Cassava Sciences, Inc. (SAVA)

A clinical-stage biotechnology company



Symbol	Name	Last Price	Change	% Change
LI	Li Auto Inc.	32.40	-0.22	-0.67%
PTPI	Petros Pharmaceuticals, Inc.	2.0900	+0.2600	+14.21%
FSR	Fisker Inc.	19.79	-0.59	-2.89%
NRXP	NRx Pharmaceuticals, Inc.	6.75	+2.25	+50.00%
BNTX	BioNTech SE	348.00	+43.24	+14.19%
KRYS	Krystal Biotech, Inc.	39.81	-2.63	-6.20%
APDN	Applied DNA Sciences, Inc.	5.30	+0.43	+8.83%
ADGI	Adagio Therapeutics, Inc.	25.12	+6.39	+34.12%
BFRI	Biofrontera Inc.	7.90	+2.68	+51.34%
MRNA	Moderna, Inc.	329.63	+56.24	+20.57%
PTN	Palatin Technologies, Inc.	0.5576	+0.0432	+8.40%
PRPO	Precipio, Inc	1.9100	-0.0400	-2.05%

Data crawling - Requirements

1 **BeautifulSoup**

Parse HTML/XML documents

2 **Pandas**

Export files

3 **Selenium**

Controls browser

Data description

1. **Date**

   Day of trading

2. **Open**

   Starting price in day

3. **Close**

   Ending price in day

4. **Adjusted close**

   Close price, adjusted after business actions

5. **High**

   Highest price in day

6. **Low**

   Lowest price in day

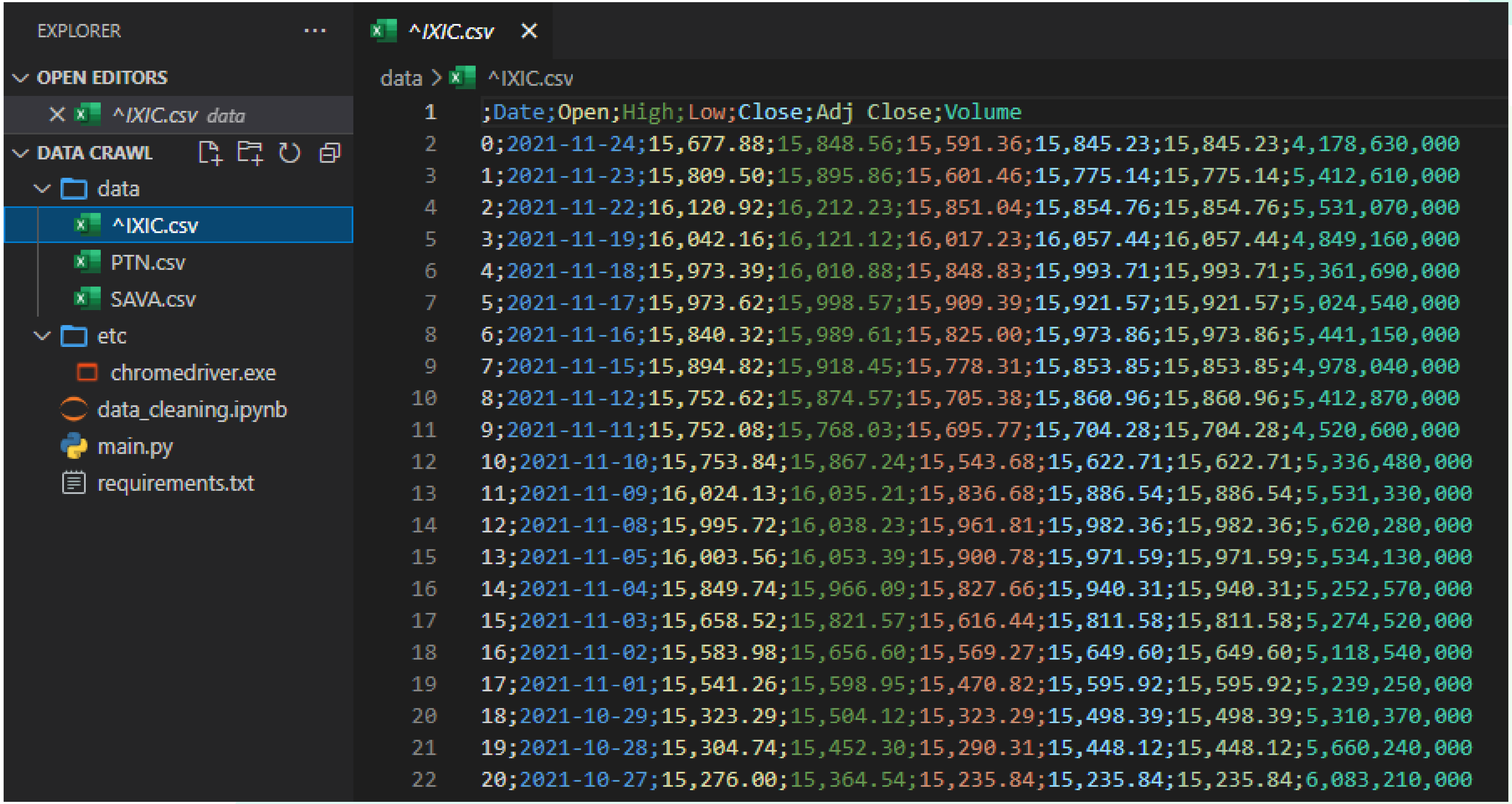
7. **Volume**

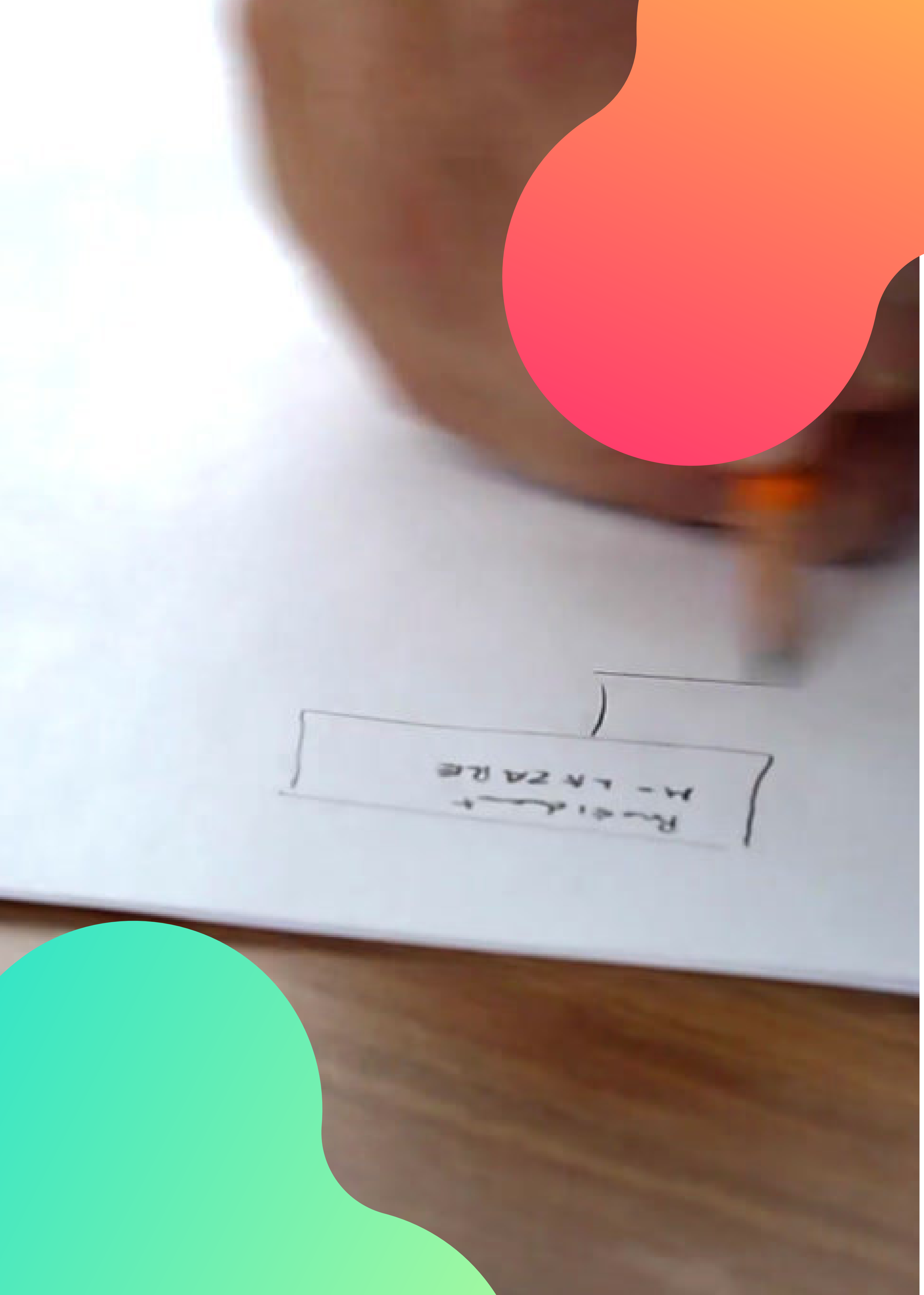
   Number of shares traded in day

Data crawling - Steps

1. Use Selenium to drive a browser, to go to source page
2. Scroll to bottom of page until no longer scrollable
3. Extract table rows from complete page
4. Export data to .csv file

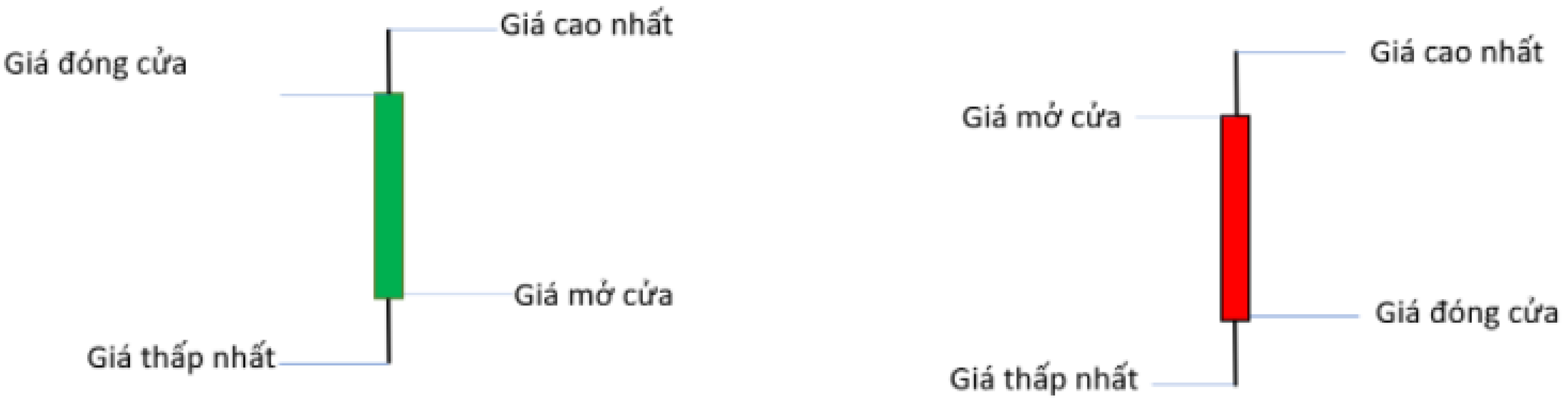
Data crawling - Results





DATA VISUALIZATION

Price volatility: Candlestick charts



A candle represents price behavior in a day

- Opening price vs. closing price
- Highest price and lowest price

Trading volume

Number of shares traded over a period of time

Sample candlestick chart analysis

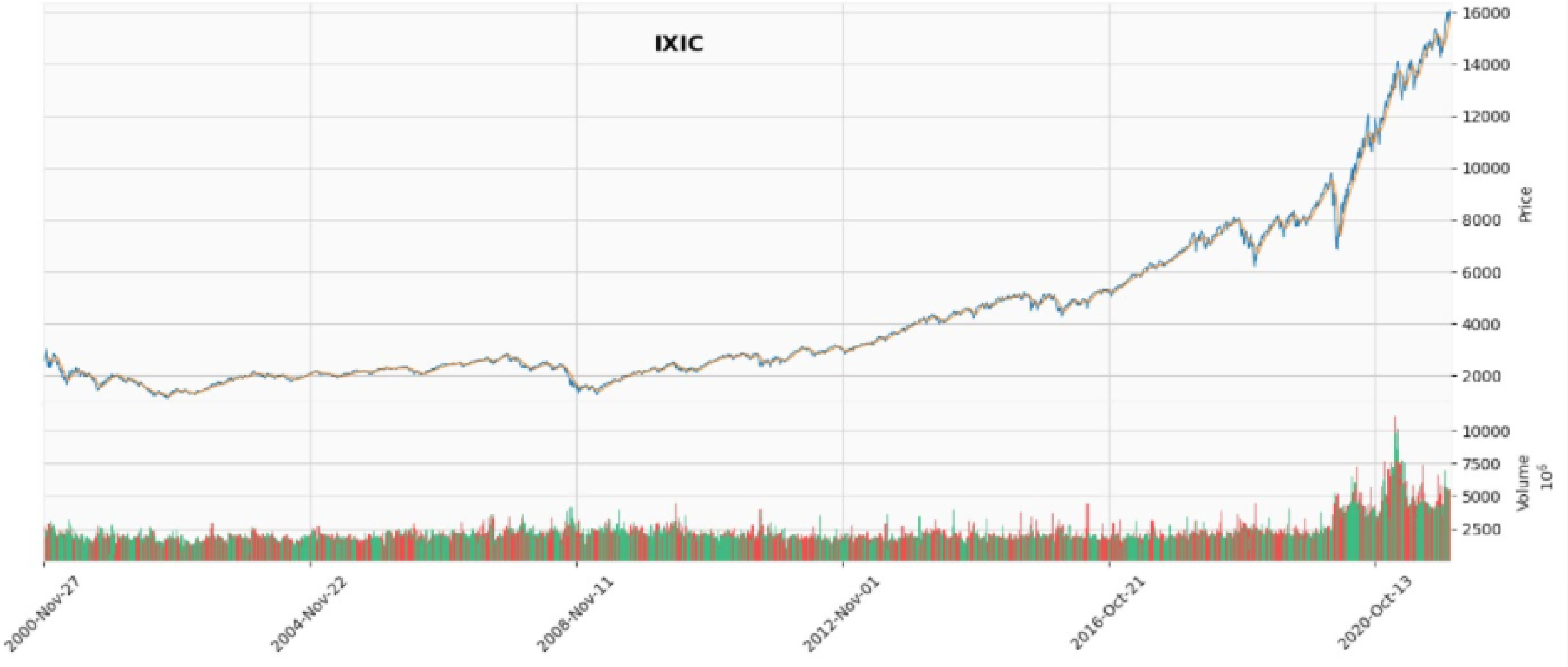
- Case A
- Case B

Overall observations (25/11/2000 - 25/11/2021)

- NASDAQ Composite (^IXIC)
- Palatin Technologies (PTN)
- Cassava Sciences, Inc. (SAVA)

NASDAQ Composite (^IXIC)

Overall observation (25/11/2000 - 25/11/2021)

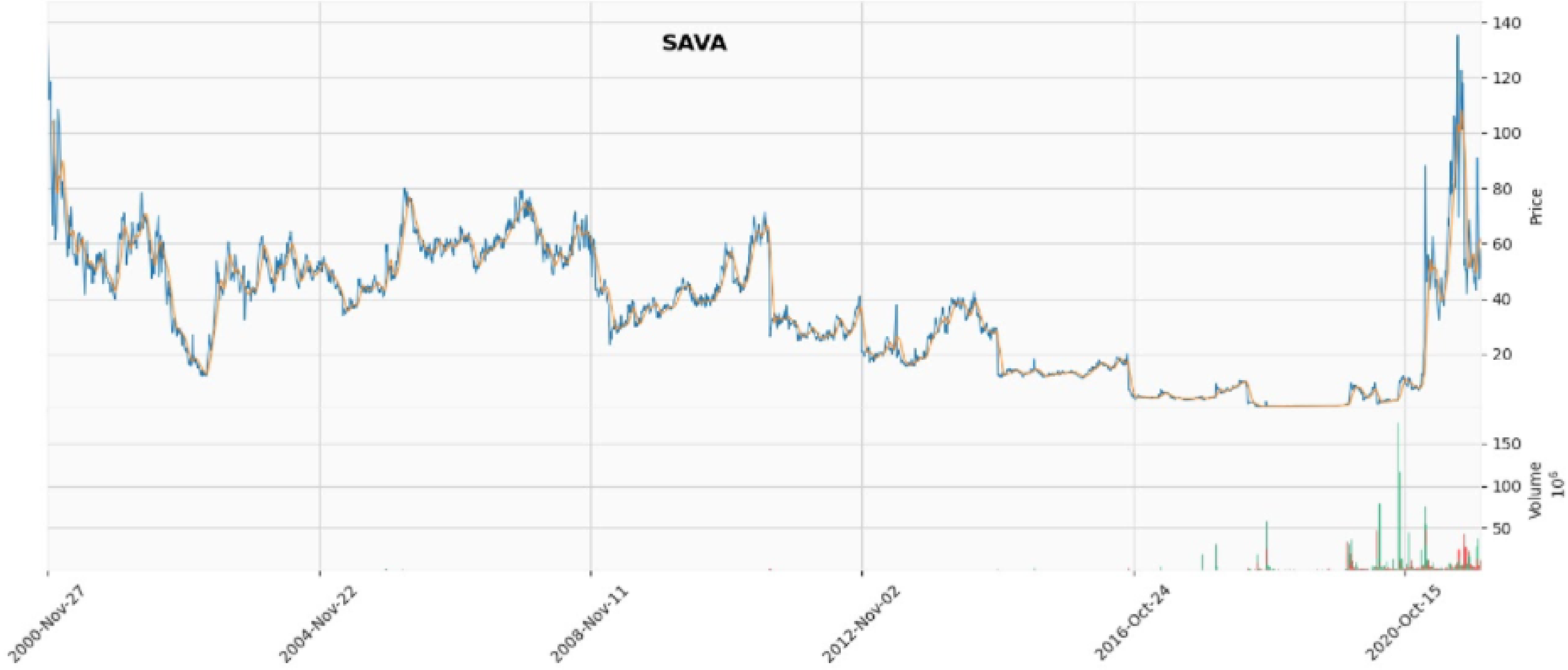


Palatin Technologies

Overall observation (25/11/2000 - 25/11/2021)

Cassava Sciences, Inc.

Overall observation (25/11/2000 - 25/11/2021)



Latest observations (3/5/2021- 25/11/2021)

- NASDAQ Composite (^IXIC)
- Palatin Technologies (PTN)
- Cassava Sciences, Inc. (SAVA)

NASDAQ Composite (^IXIC)

Latest observation (3/5/2021- 25/11/2021)

NASDAQ Composite (^IXIC)

Latest observation (3/5/2021- 25/11/2021)

Palatin Technologies

Latest observation (3/5/2021- 25/11/2021)

Palatin Technologies

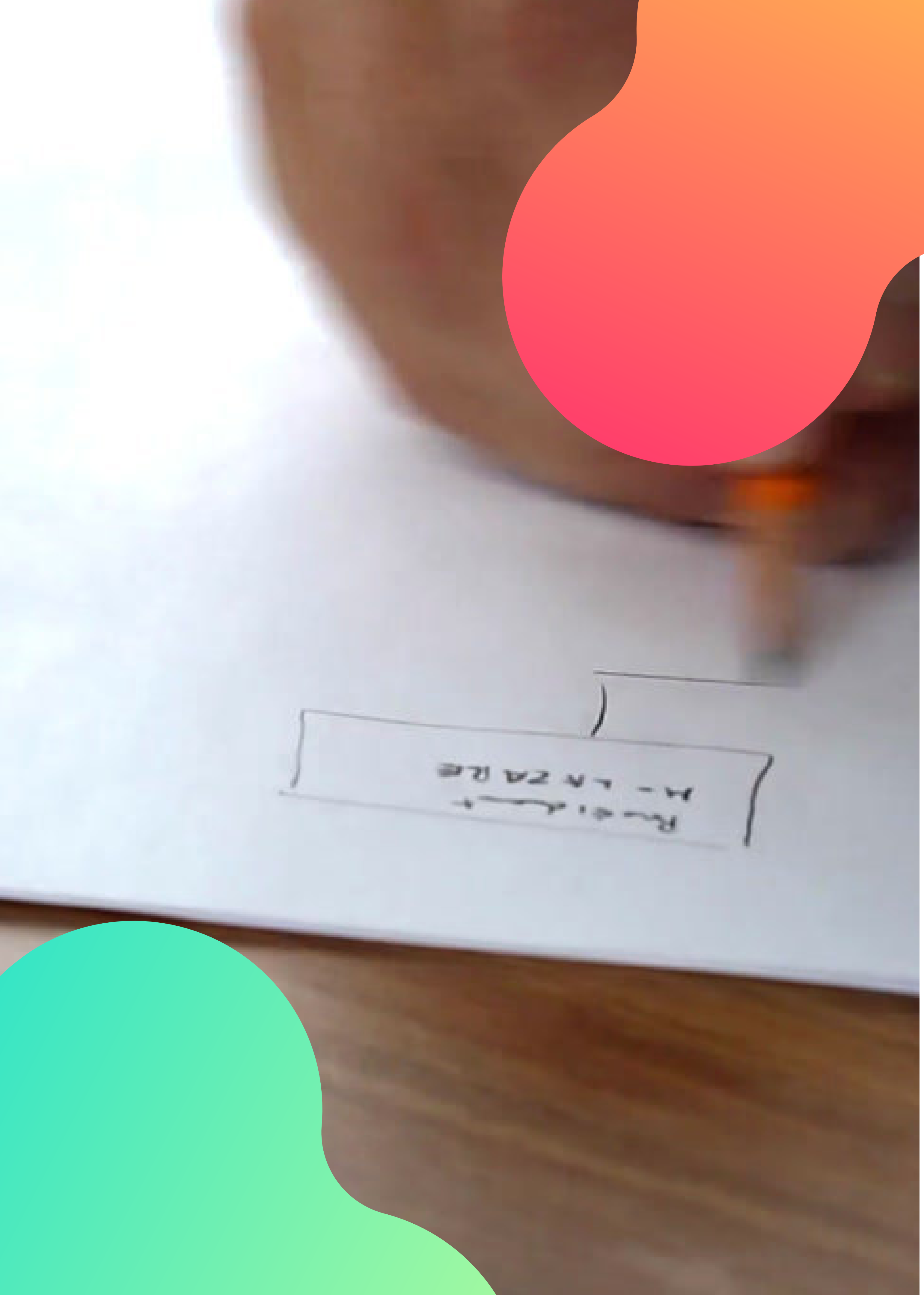
Latest observation (3/5/2021- 25/11/2021)

Cassava Sciences, Inc.

Latest observation (3/5/2021- 25/11/2021)

Cassava Sciences, Inc.

Latest observation (3/5/2021- 25/11/2021)



METHOD

Overview: Gaussian Process

A Gaussian process is a generalization of the Gaussian distribution - it represents a probability distribution over *functions* which is entirely specified by a mean and covariance *functions*. Mathematical definition would be then as follows:

Definition: *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

Let x be some process $f(x)$. We write:

$$f(x) \sim GP(m(\cdot), k(\cdot, \cdot)),$$

where $m(\cdot)$ and $k(\cdot, \cdot)$ are the mean and covariance functions, respectively:

$$\begin{aligned} m(x) &= E[f(x)] \\ k(x_1, x_2) &= E[(f(x_1) - m(x_1))(f(x_2) - m(x_2))]. \end{aligned}$$

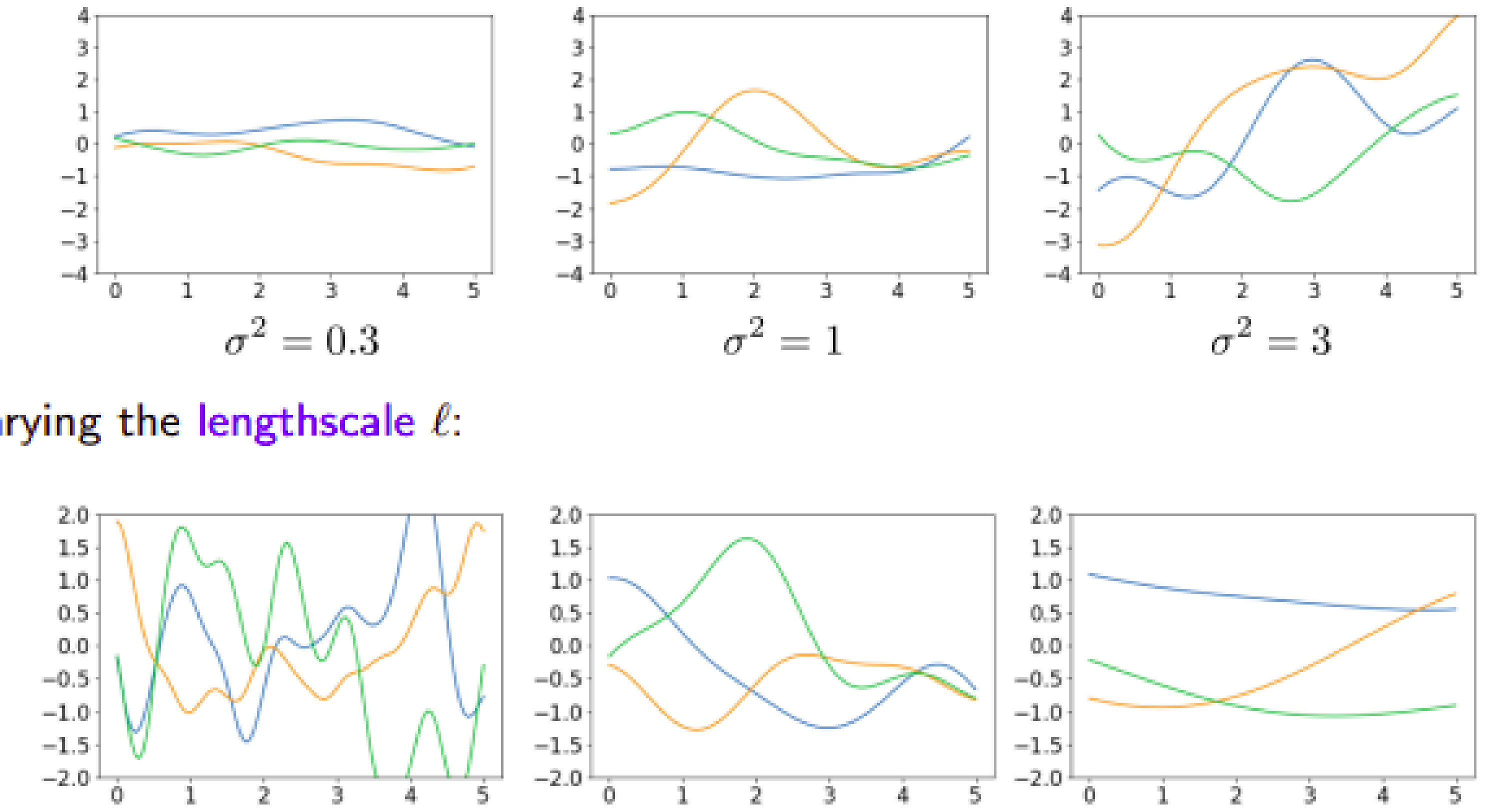
Hyperparameters: GP Kernel

$$k_{\mathrm{SE}}(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2\ell^2}\right)$$

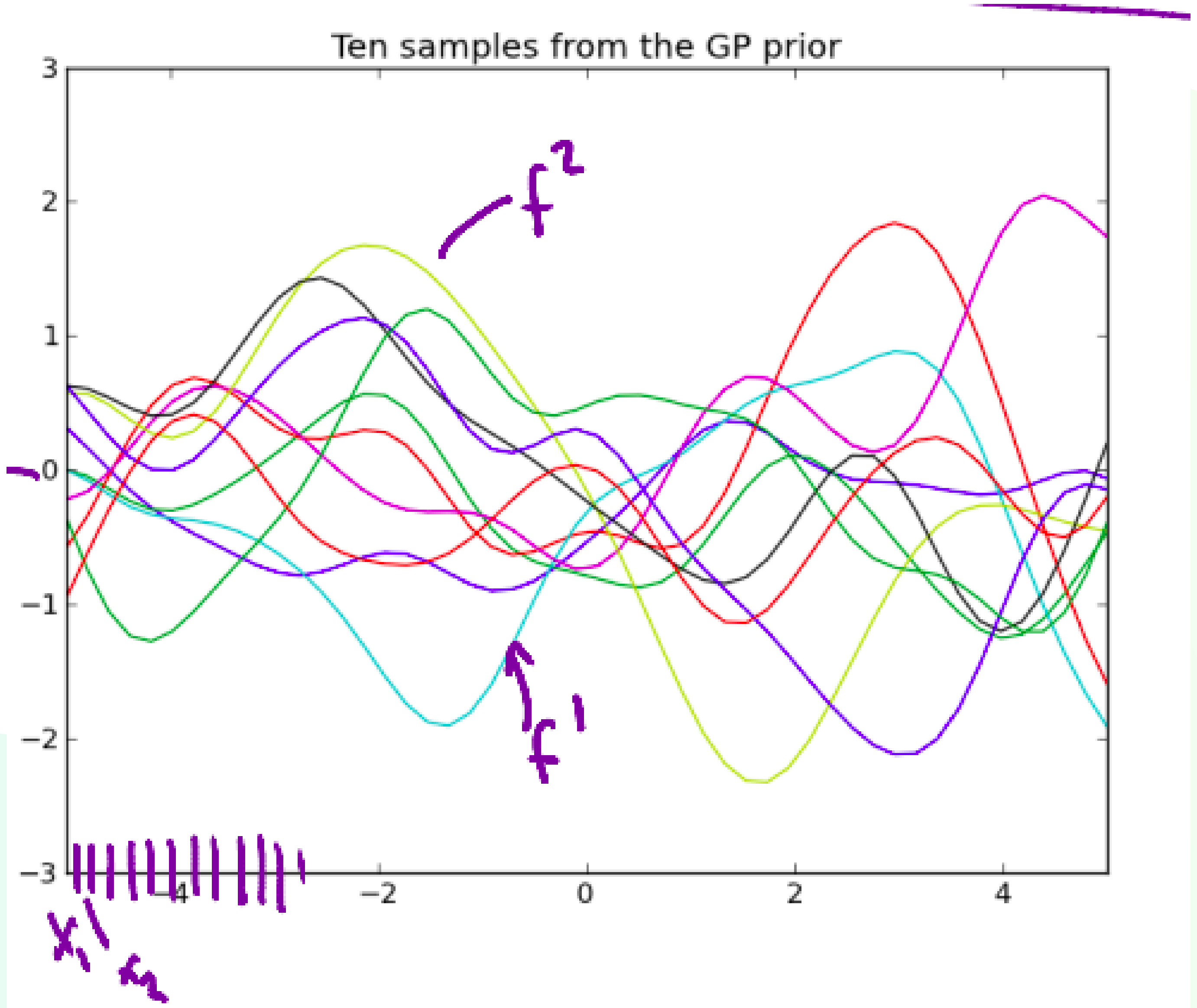
- The hyperparameters determine key properties of the function.
- Varying the output variance σ^2:

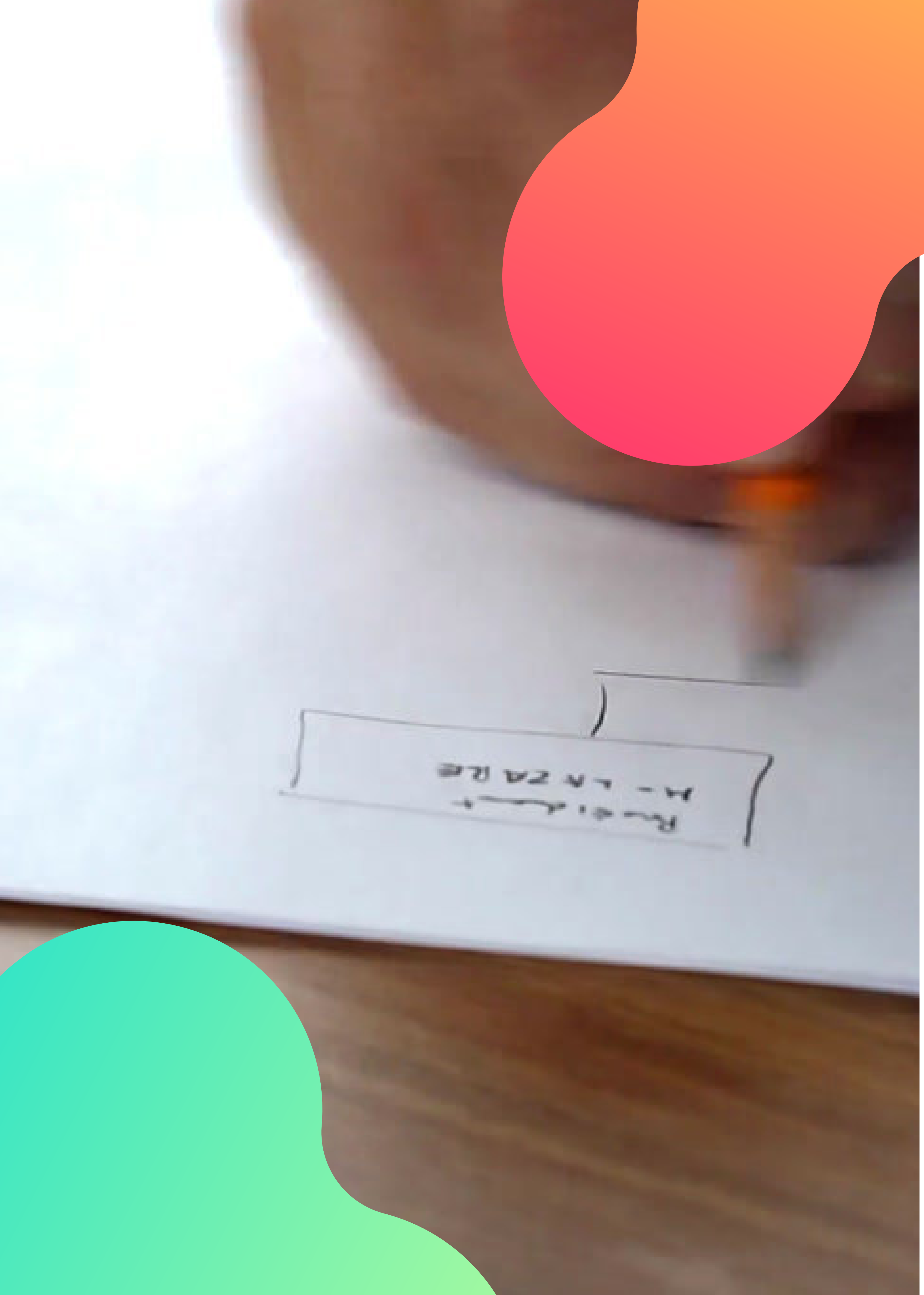
$\sigma^2 = 0.3$ $\sigma^2 = 1$ $\sigma^2 = 3$

- Varying the lengthscale ℓ:



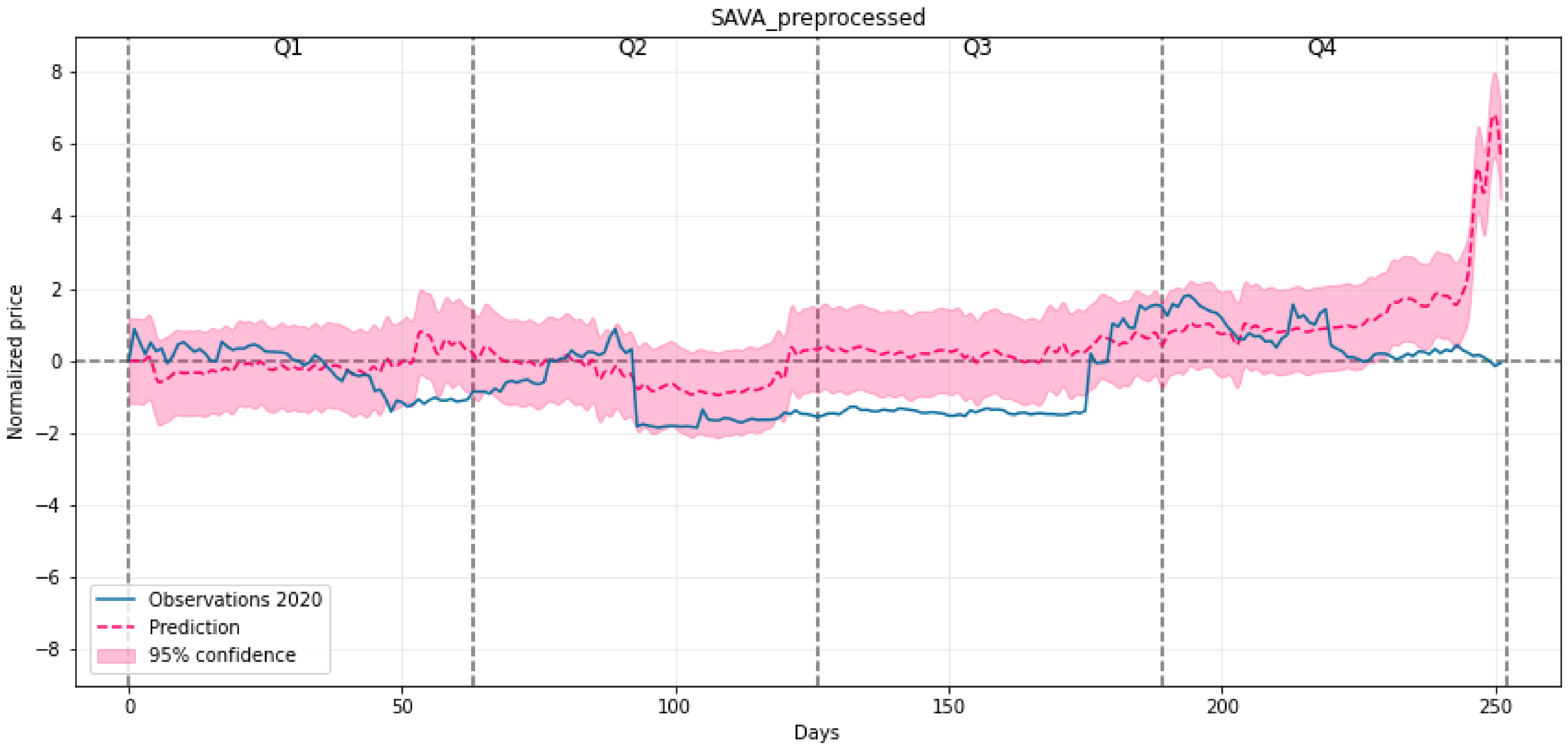
$\ell = 0.3$ $\ell = 1$ $\ell = 3$



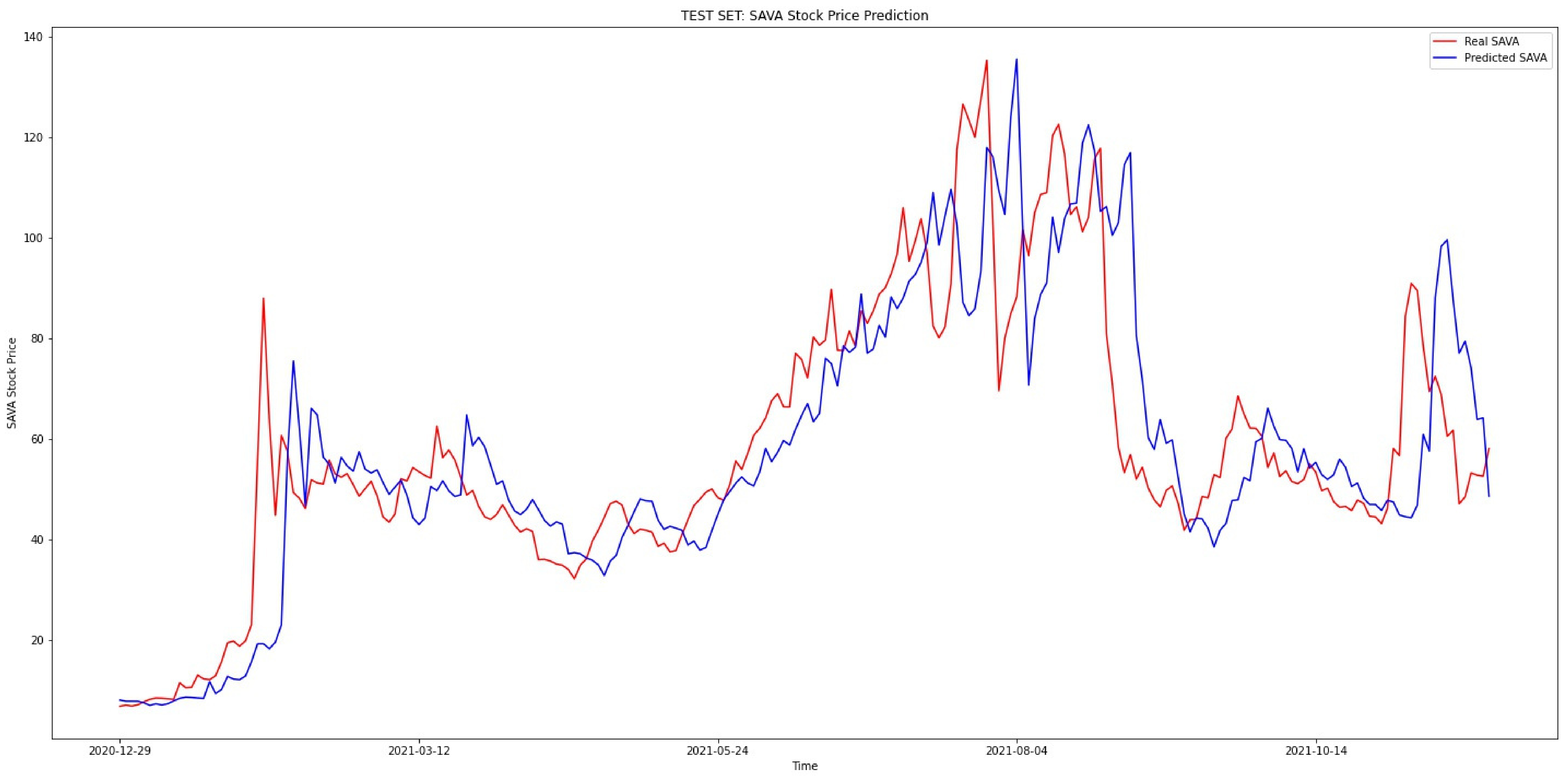


EVALUATION

Result on Gaussian Process: SAVA

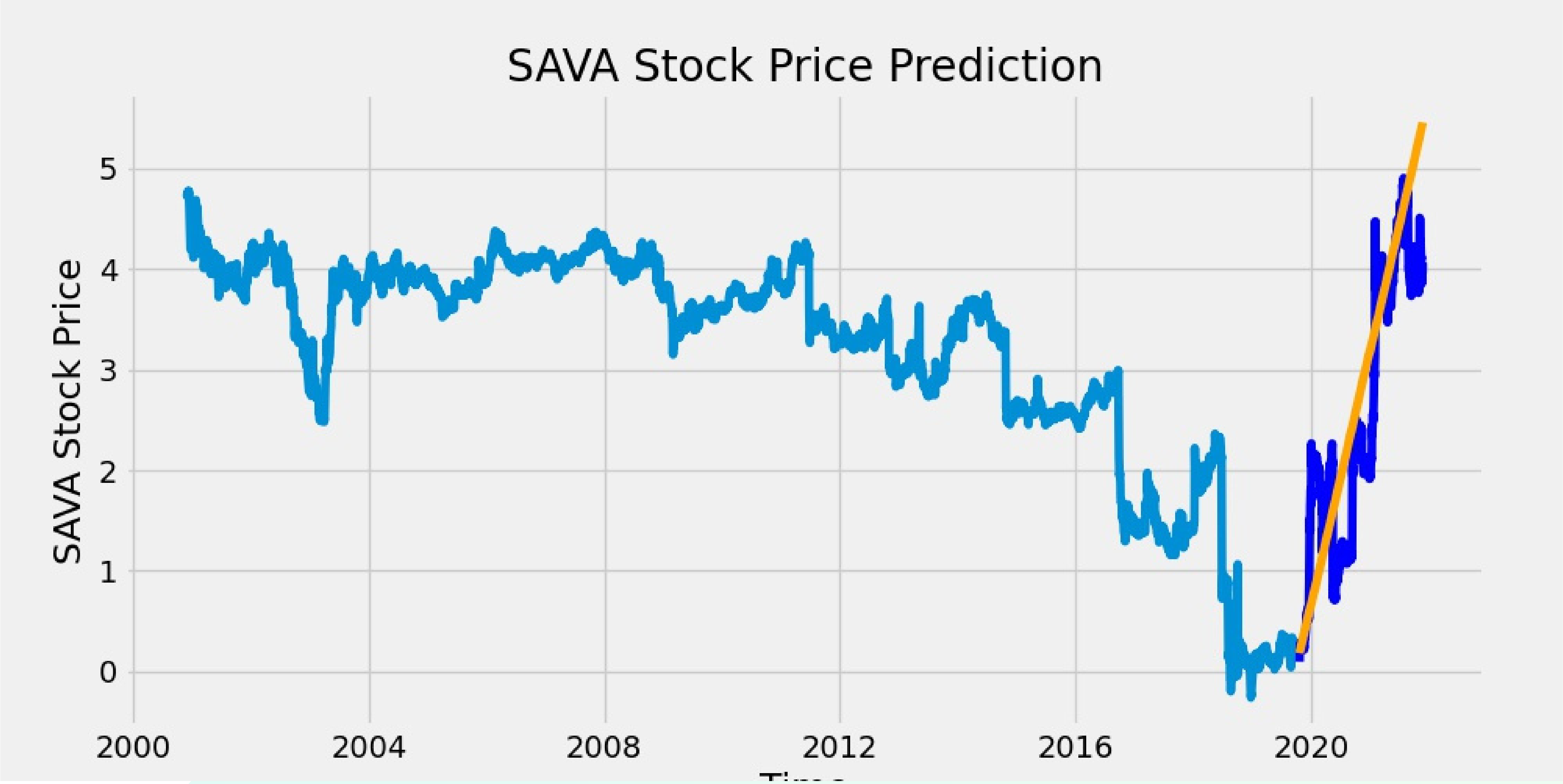


Result on other methods: LSTM



Result on other methods: Linear Regression

Result on other methods: ARIMA



Comparison

Type	Gaussian Process	LSTM	Linear Regression	ARIMA model
MSE	152.5147257	28.7624804	282.5947357	0.56767490829
MAE	8.15730316	4.172498876	12.29430316	0.610451537752